Automatic Classification of Russian Texts for Didactic Purposes

Batinić, Dolores ; Birzer, Sandra ; Zinsmeister, Heike

Published in: Trudy meždunarodnoj konferencii „Korpusnaja lingvistika - 2017". 27-30 ijunja 2017 g., Sankt-Peterburg, 2017. Pp. 9-15

D. Batinić, S. Birzer, H. Zinsmeister

AUTOMATIC CLASSIFICATION OF RUSSIAN TEXTS FOR DIDACTIC PURPOSES

Abstract. In this paper we present the results of an automatic classification of Russian texts into three levels of difficulty. Our aim is to build a study corpus of Russian, in which a L2 student is able to select texts of a desired complexity. We are building on a pilot study, in which we classified Russian texts into two levels of difficulty. In the current paper, we apply the classification to an extended corpus of 577 labelled texts. The best-performing combination of features achieves an accuracy of 0,74 within at most one level difference.

Keywords. L2 Russian, didactic corpus, text complexity, text classification.

1. Introduction

Working with linguistic corpora is an integral part of many foreign language studies (e.g. [Römer 2008; Steinbach & Birzer 2012]). Analyzing texts which are beyond the learner's level may frustrate them and hinder the learning process, whereas reading texts beneath their proficiency may impede their improvement. We argue that the possibility of being able to select a desired level of text difficulty will bring benefits to L2 corpus users in their learning experience. Our goal is to create a Levelled Study Corpus (LeStCor) for L2 learners of Russian that involves filtering options for different complexity levels and a didactic highlighting of difficult morphosyntactic structures [Birzer & Zinsmeister 2016]. In a pilot study of automatic two-level classification on 209 texts, we obtained satisfactory results by considering both surface-oriented features adopted from general readability assessments and more linguistically informed features [Batinić et al. 2016]. In the current paper, we apply a modified classification model to an extended training corpus. In order to discriminate between the difficulty levels, we train an NLTK Naive Bayes classifier on manually labelled texts.

2. Related work

The assessment of text difficulty for native speakers has its origins in the 1920s. Surface-oriented readability measures allowed the researchers to compare different texts in an objective way. More recent approaches integrate features that address a) the lexical coverage of a text, b) parts of speech, c) syntactic structures, e) crosssentential features like the referential overlap and f) relations between clauses triggered by discourse connectives [Benjamin 2012]. Studies aimed at the difficulty level for L2 learners have the underlying hypothesis that L2 learners perceive text comprehensibility dif-

ferently than L1 students [François 2014]. [Chinkina & Meurers 2016], for example, integrated 87 linguistic features to classify English texts into three difficult levels for L2 learners. [Baranova, & Elipašheva 2014] developed a rule-based tool for analyzing the difficulty of Russian texts. Machine learning approaches exploit the strength of different features in a data-driven probabilistic way (e. g. [Xia et al. 2016]). Our work is similar in approach to [Curto et al. 2015], who studied an automatic five- and three-level classification of a small set of Portuguese texts.

3. Material

We selected 577 texts originating from the Test of Russian as a Foreign Language (TORFL, Russian: TRKI) reading and listening tasks. We also included newspaper articles from Ria Novosti[1] and labelled them as the most advanced level (Class III). The number of texts was distributed similarly across classes. A detailed view of the corpus stratification can be found in Table 1. We added the corresponding levels of the Common European Framework of References for Languages (CEFR) for comparison.

Table 1. **TRKI proficiency levels and sampling of the corpus**

Class	TRKI	CEFR	Sem	Texts	Texts/class
I	elementary	A1	1st	57	180
	basis	A2	1st	123	
II	1	B1	2nd	109	206
	2	B2	3rd	97	
II	3	C1	4th	52	191
	4	C2	indep.	25	
		C3	indep.	114	

In order to be able to apply diverse lexical and morphosyntactic features, all texts were tagged and lemmatized with TreeTagger using Russian parameter files, trained on the disambiguated version of the Russian National.[2]

[1] https://ria.ru/ (05.04.2017).

[2] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (05.04.2017).

4. Feature selection

We assumed that the most indicative feature of text difficulty consists of the proportion of basic vocabulary in texts. In order to operationalize the basic vocabulary we used vocabulary lists originating from the textbooks Dialog 1 and 2,[3] which correspond to base and elementary level of language proficiency (A1 and A2 according to CEFR). After preprocessing, the list of basic vocabulary contained 1144 lemmas. We extended the list of basic vocabulary with the list of the 5000 most frequent Russian lemmas compiled by [Sharoff 2002], which proved to be a good text difficulty predictor in our previous study. In addition, we also considered numerals, named entities, pronouns and internationalisms (gathered from Wikipedia's list of internationalisms in Russian), since they are also easily understandable to a L2 student, although not (necessarily) provided in the vocabulary or frequency lists.

With regard to other features, we measured the average number of adverbial participles, perfect participles, parts of speech (nouns, verbs, pronouns, adjectives, adverbs, adpositions, conjunctions, and particles) and abstract words per sentence. Knowing that morphosyntactic features such as participles are introduced at the intermediate proficiency levels (TRKI 1 and TRKI 2), we expected them to be highly discriminative. In order to approximate the number of abstracta in texts, we counted Russian words ending with *-изм* '*-ism*', *-остьь* '*-ness*', *-ство* '*-ship*', *-ома* '*-ness*', *-ание / -ение* (nominalized verbs). We also experimented with other features (lexical density, type/token), which, however, did not prove to be sufficiently informative.

We set the Flesch-Kincaid score adapted to Russian [Oborneva 2016] as our baseline. Flesch-Kincaid approximates the readability of a text by taking into account surface features such as the number of words, sentences and syllables in a text.

5. Results and Discussion

We performed a classification with Naive Bayes (NLTK)[4], and 10-fold cross validation. The values for all the features were set heuristically and by considering the distributions in Figure 1. The highest accuracy (0,74) was achieved by combining the features common words, abstract words, past

[3] Dialog. Lehrwerk für den Russischunterricht. Neue Generation. Bd. 1/2. (2016/2017) [Dialogue. Textbook for Russian language instruction. New Generation. Vol. 1/2.] Berlin: Cornelsen.

[4] www.nltk.org (05.04.2017).

participle and adverbial participle (thresholds: ≥ 95%, < 89% and < 85% for common words, < 0,50, > 1,30 and > 3 for abstracta, > 0 for adverbial participles, > 0 for past participles and > 0,40 for both participles together).

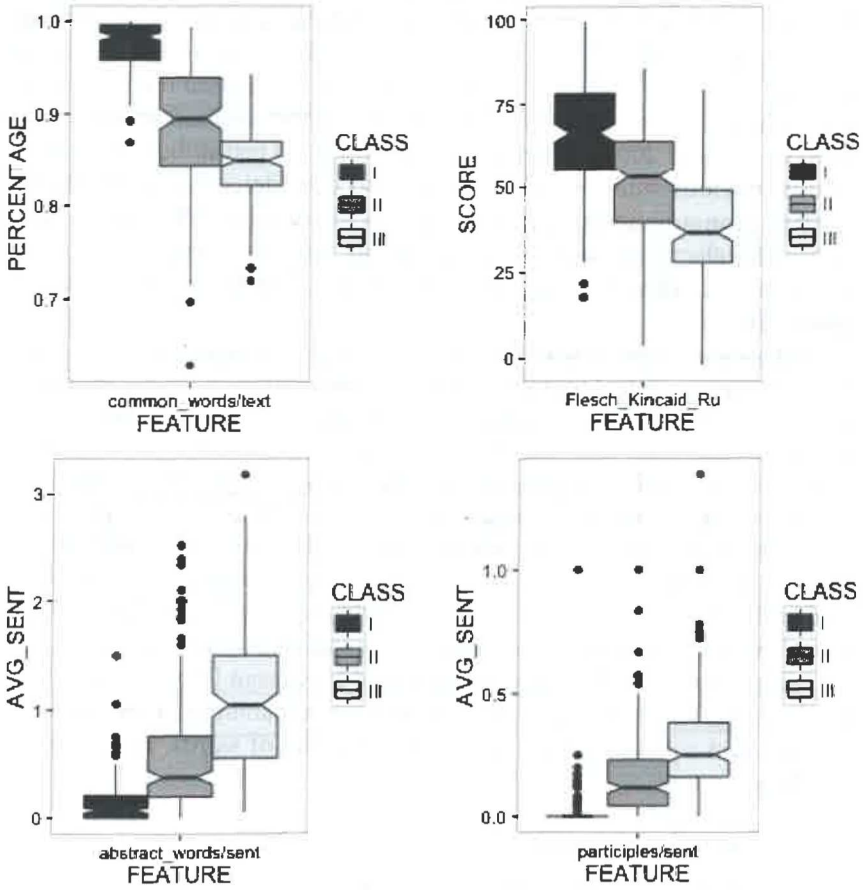

Figure 1. **Boxplots of feature distributions across classes**

As expected, the proportion of common words proved to be the most informative feature: with this feature alone the accuracy rose to 0,68. With an accuracy of 0,63, the combination of average numbers of adverbial and past participle also confirmed the assumption of being good predictors for a three-level text classification. The baseline accuracy of 0,50, which

was the highest achieved by Flesch Kincaid (threshold: > 60) was hence outperformed in a significant manner.

With the best performing combination of features, the classifier only missed within at most one level difference in all ten test sets. The erroneously predicted levels are in many cases also those, whose levels may be disputable even by human judgments.

The feature *common words* proved to be highly informative, especially for discriminating between Class I and Class II. However, differentiating between Class II and Class III based solely on vocabulary lists appears more demanding, given the fact that the vocabulary threshold between intermediate and proficient learners is difficult to estimate. Vocabulary acquisition on a high intermediate level is likely to vary from student to student and may depend on the field in which they choose to intensify their L2 study. Hence, for discriminating between intermediate and advanced levels it might be more appropriate to continue to rely on morphosyntactic features instead of gathering other vocabulary lists. It may as well be advantageous to introduce new features, such as multiword expressions or syntactic formulae, with which a proficient learner should be familiar.

As much as a readability measure such as Flesch-Kincaid may be considered to be a useful indicator of reading comprehension for both native speakers and L2 learners, one must not rely on it entirely when selecting appropriate texts for language learning purposes. The readability score does not in fact measure the level of text difficulty in terms of linguistic features, which prove to be well suited for text classification directed to L2 learners. Measures that only rely on surface features may easily fail in texts with a dialog-like structure, in which the sentences may be short, but the vocabulary may be exigent. On the contrary, passages that may appear unreadable because of long words and sentences might be easily understood by an adult L2 learner if the vocabulary and syntactic structures are familiar.

6. Conclusion

We conducted an automatic classification of Russian texts into three levels of difficulty. The classifier achieved an accuracy of 0,74 with the best predictors consisting of lexical and morphosyntactic features. In a future study, we aim to extend the set of features in order to consider different syntactic and multiword phenomena.

References

1. *Baranova J., Elipašheva T.* (2014), Sozdanie vspomogatel' nogo informacionnogo resursa dlja analiza učebnych tekstov narusskom jazyke. [Creating an auxiliary information resource for the analysis of Russian as a Foreign Language reading texts.]. In: Anis'kin N. (ed.), Čelovek v informacionnom prostranstve. Jar., JaGPU, pp. 232–246.
2. *Batinić D., Birzer S., Zinsmeister H.* (2016), Creating an extensible, levelled study corpus of Russian. Proceedings of KONVENS 2016, Bochum, pp. 38–43.
3. *Benjamin R.* (2012), Reconstructing readability. In: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24, pp. 63–88.
4. *Birzer S., Zinsmeister H.* (2016), The utility of colour markup in texts for learners of Russian. In: Proceedings of the 8th International Biannual Conference "Applied Linguistics in Research and Education". St Petersburg, pp. 139–145.
5. *Chinkina M., Meurers D.* (2016), Linguistically Aware Information Retrieval. In: Providing Input Enrichment for Second Language Learners. Proceedings of BEA. San Diego, pp. 188–198.
6. *Curto P., Mamede N., Baptista B.* (2015), Automatic text difficulty classifier. In: Assisting the selection of adequate reading materials for European Portuguese teaching. Proceedings of CSEDU 2015, Lisboa, pp. 36–44.
7. *François T.* (2014), An analysis of a French as a Foreign Language corpus for readability assessment. In: Proceedings of NEALT, Linköping, pp. 13–32.
8. *Oborneva I. V.* (2006), Avtomatizirovannaja ocenka složnosti učebnych tekstov na osnove statističeskich parametrov. Avtoreferat dissertacii na soiskanie učënoj stepeni kandidata pedagogičeskich nauk. [Automatic rating of the degree of difficulty of textbook texts on the bases of statistical parameters.] Available at: http://naukapedagogika.com/pedagogika-13-00-02/dissertaciyaavtomatizirovannaya-otsenka-slozhnosti-uchebnyh-tekstov-naosnove-statisticheskih-parametrov (05.04.2017).
9. *Römer U.* (2008), Corpora and language teaching. In: Lüdeling A., Kytö M. (eds.), Corpus Linguistics. An International Handbook (vol. 1). Berlin, pp. 112–130.
10. *Sharoff S.* (2002), Meaning as use. In: Exploitation of aligned corpora for the contrastive study of lexical semantics. Proceedings of LREC 2002, Las Palmas, pp. 447–452.
11. *Steinbach A., Birzer S.* (2011), Authentisches Sprachmaterial schnell gefunden. Das Potenzial russischer Textkorpora im Russischunterricht [Authentic texts at your disposal: the potential of text corpora for Russian as a Foreign Language classes]. In: Praxis Fremdsprachenunterricht, 2, pp. 7–10.
12. *Xia M., Kochmar E., Briscoe T.* (2016), Text readability assessment for second language learners. Proceedings of BEA'16, San Diego, pp. 12–22.

Dolores Batinić
Institut für Deutsche Sprache (Germany)
E-mail: batinic@ids-mannheim.de

Sandra Birzer
University of Innsbruck (Austria)
E-mail: Sandra.Birzer@uibk.ac.at

Heike Zinsmeister
Universität Hamburg (Germany)
E-mail: heike.zinsmeister@uni-hamburg.de